UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/573,482	03/24/2006	David Patterson	27309U	2569

20529 7590 09/15/2008

NATH & ASSOCIATES
112 South West Street
Alexandria, VA 22314

EXAMINER	
RUIZ, ANGELICA	
ART UNIT	PAPER NUMBER
2169	
MAIL DATE	DELIVERY MODE
09/15/2008	PAPER

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Office Action Summary	Application No.	Applicant(s)	
	10/573,482	PATTERSON ET AL.	
	Examiner	Art Unit	
	ANGELICA RUIZ	2169	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE *3* MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☒ Responsive to communication(s) filed on *27 May 2008*.

2a)☒ This action is **FINAL**. 2b)☐ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) *1-18* is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) *1-18* is/are rejected.

7)☐ Claim(s) _____ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☐ The specification is objected to by the Examiner.

10)☐ The drawing(s) filed on _____ is/are: a)☐ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☒ Notice of References Cited (PTO-892)
2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
3) ☐ Information Disclosure Statement(s) (PTO/SB/08) Paper No(s)/Mail Date _____.
4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____ .
5) ☐ Notice of Informal Patent Application
6) ☐ Other: _____.

DETAILED ACTION

1. The Action is responsive to Applicant's amendment, filed on May 27, 2008.

2. It is acknowledged that as a result of the amendment, Claims 1, 4, 13, 14, 15, 17, and 18 have been amended.

3. Claims 1-18 are pending.

Response to Arguments

4. Applicant's arguments with respect to claims 1-18 have been considered but are moot in view of the new grounds of rejection necessitated by Applicant's amendment of the claims.

Applicant argues in substance that Sahami discloses clustering structured data, and Claim 1 relates to identifying cluster attractors for documents comprising unstructured text, Examiner disagrees with applicant because there is no such claim of "unstructured text" in the mentioned claim language.

The claims and only the claims form the metes and bounds of the invention. "Though understanding the claim language may be aided by explanations contained in the written description, it is important not to import into a claim limitations that are not part of the claim. For example, a particular embodiment appearing in the written description may not be read into a claim when the claim language is broader than the embodiment." Superguide Corp. v. DirecTV Enterprises, Inc., 358 F.3d 870, 875, 69 USPQ2d 1865, 1868 (Fed. Cir. 2004). See also Liebel-Flarsheim Co. v. Medrad Inc., 358 F.3d 898, 906, 69 USPQ2d 1801, 1807 (Fed. Cir. 2004)

Claim Objections

5. Claims 1 and 13 are objected due to being Duplicated Claims.

Applicant is advised that should claim [1] be found allowable, claims [13] will be objected to under 37 CFR 1.75 as being a substantial duplicate thereof. When two claims in an application are duplicates or else **are so close in content that they both cover the same thing, despite a slight difference in wording**, it is proper after allowing one claim to object to the other as being a substantial duplicate of the allowed claim. See MPEP § 706.03(k).

Claim Rejections - 35 USC § 101

6. The text of those sections of Title 35, U.S. Code not included in this action can be found in a prior office action.

7. In view of the amendments to the claims, the Examiner withdraws all pending rejections under 35 U.S.C. 101.

Claim Rejections - 35 USC § 103

8. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

9. Claims 1-18 are rejected under 35 U.S.C. 103(a) as being unpatentable over **Sahami et al (US Publication No. 2003/0065635 A1)**, in view of **Tukey et al (US Patent No. 5,787,422)**.

> **- A method of determining cluster attractors for a plurality of documents, each document comprising at least one term,** each term comprising one or more words**, the method comprising: calculating, in respect of each term, a probability distribution indicative of the frequency of occurrence of the, or each, other term that co-occurs with said term in at least one of said documents; calculating, in respect of each term, the entropy of the respective probability distribution; selecting at least one of said probability distributions as a cluster attractor depending on the respective entropy value.**

(Abstract, "…include methods for identifying clusters in a database, data warehouse or data mart. The identified clusters can be meaningfully understood by a list of the attributes and corresponding values for each of the clusters…") and (Par [0058], "The do not care entry is useful in probabilistic algorithms because the frequency of different attribute values is easily computed. For example, the probability that x.sub.1=a is: 1 P (x 1 = a) = (x 1 = a , x 2 = * , x 3 = *) (x 1 = * , x 2 = * , x 3 = *) ,") and (Par [0059]) and (Par [0062], "The mutual…") and (Par [0027], "The previously discussed techniques were all oriented towards clustering entire sets of data. COBWEB is an online, or incremental approach to clustering. FIG. 4 shows a COBWEB tree structure with

clusters. The clusters are the nodes of the tree. FIG. 4 shows a new data point, X, to be added to the data set. COBWEB is based on **a probability distribution** …").

However Sahami, does not specifically disclose:

- ***cluster attractors***

- each term comprising one or more words

On the other hand Tukey discloses the above claimed features as follow:

(Abstract and Claim 1, "…identifying an attractor for each of a plurality of clusters…"). (Col. 6 and Col. 7, lines 65-67 and 1-9, respectively, "…A feature may be a word, a statistical phrase…").

Therefore, it would have been obvious to a person of ordinary skill in the art at the time of invention was made to incorporate the teachings of Tukey into the method of Sahami to include cluster attractors based on the probability distribution and entropy value. The modification would have been obvious because one of the ordinary skills in the art would implement an effective approach to find the best similarities using the cluster attractors and get the closest match to the comparison between the set of values to give the user the best probable result.

As per Claim 2, the rejection of Claim 1 is incorporated and further Sahami discloses:

> **- wherein each probability distribution comprises, in respect of each co-occurring term, an indicator that is indicative of the total number of**

instances of the respective co-occurring term in all of the documents in which the respective co-occurring term co-occurs with the term in respect of which the probability distribution is calculated.

(Par [0012], “One common technique for identifying clusters is the k-means technique. … FIG. 1 shows a set of clusters defined by centroids through the k-means technique. The data points are indicated with "." in the two dimensional data domain space. The centroids are indicated with "x". The resulting clusters are formed by those data points within a certain distance of the centroids as indicated by the ellipsoids.) and (Par [0015], "The k-means technique is also fairly computationally expensive, especially given that additional computational resources will have to be used if any analysis of the clusters is required. In big-O notation, the kmeans algorithm is O(knd), where k is the number of centroids, n is the number of data points, and d is the number of iterations.”).

As per Claim 3, the rejection of Claim 1 is incorporated and further Sahami discloses:

- wherein each probability distribution comprises, in respect of each co-occurring term, an indicator comprising a conditional probability of the occurrence of the respective co-occurring term in a document given the appearance in said document of the term in respect of which the probability distribution is calculated.

(Par [0015], “The k-means technique is also fairly computationally expensive, especially given that additional computational resources will have to be used if any analysis of the

clusters is required. In big-O notation, the kmeans algorithm is O(knd), where k is the number of **centroids**, n is the number of data points, and d is the number of iterations.") and (Par [0066], "As subsequent clusters are defined, the mutual information will be computed as a **conditional probability** based on the clusters that have already been identified: MI(x.sub.i,x.sub.j.vertline.Z), where Z is the set of features previously split on, e.g. Z={x.sub.m=a, x.sub.n=d}.").

As per Claim 4, the rejection of Claim 1 is incorporated and further Sahami discloses:

> **- wherein each indicator is <u>normalized with respect to the total number of terms in the, or each, document in which the term</u> in respect of which the probability distribution is calculated appears.**

(Par [0027], "The previously discussed techniques were all oriented towards clustering entire sets of data. COBWEB is an online, or incremental approach to clustering. FIG. 4 shows a COBWEB tree structure with clusters. The clusters are the nodes of the tree. FIG. 4 shows a new data point, X, to be added to the data set. COBWEB is based on **a probability distribution** …").

However Sahami does not disclose the above underlined claimed features

On the other hand Tukey discloses the above claimed feature as follows:

(Col. 2, lines 28-47, "…consider the degree of word overlap between the two documents of interest, described as sets of words, often with frequency information.

These sets are typically represented as sparse vectors of length equal to the number of unique words (or types) in the corpus. If a word occurs in a document, its location in this vector is occupied by some positive value (one if only presence/absence information is considered, or some function of its frequency within that document if frequency is considered). If a word does not occur in a document, its location in this vector is occupied by zero. A popular similarity measure, the cosine measure, determines the cosine of the angle between two sparse vectors. If both document vectors are normalized to unit length, this is of course, simply the inner product of the two vectors. Other measures include the Dice and Jaccard coefficient, which are **normalized** word overlap counts.").

Therefore, it would have been obvious to a person of ordinary skill in the art at the time of invention was made to incorporate the teachings of Tukey into the method of Sahami to normalize with respect to the total number of items. The modification would have been obvious because one of the ordinary skills in the art would implement normalization to have more flexibility in tuning the results in the probability calculations.

As per Claim 5, the rejection of Claim 1 is incorporated and further Sahami discloses:

> **- comprising assigning each term to one of a plurality of subsets of terms depending on the frequency of occurrence of the term; and selecting, as a cluster attractor, the respective probability distribution of one or more terms from each subset of terms.**

(Abstract and Claim 1, "A method for determining a cluster from a set of data, the cluster comprised of a **subset of the set of data**, the method comprising: determining a set of attributes from the set of data, each of the set of attributes having a corresponding set of attribute values…").

However Sahami, does not specifically discloses the "**and selecting, as a cluster attractor"**

On the other hand Tukey discloses the above claimed features as follow:

(Abstract and Claim 1, **"**A method, operating in a digital computer, for searching a corpus of documents, comprising the steps of: preparing an initial structuring of the corpus into a plurality of primary overlapping clusters, wherein at least two of the plurality of primary overlapping clusters contain a single document, wherein the step of preparing an initial structuring of the corpus includes the steps of (a) identifying an attractor for each of a plurality of clusters…").

As per Claim 6, the rejection of Claim 5 is incorporated and further Sahami discloses:

- **wherein each term is assigned to a subset depending on the number documents of the corpus in which the respective term appears.**

(Abstract and Claim 1, "A method for determining a cluster from a set of data, the cluster comprised of a **subset of the set of data**, the method comprising: determining a set of attributes from the set of data, each of the set of attributes having a corresponding set of attribute values; computing a frequency information for the set of

attributes; computing a set of relation values using the frequency information, each relation value of the set of relation values…").

As per Claim 7, the rejection of Claim 5 is incorporated and further Sahami discloses:

> **- wherein an entropy threshold is assigned to each subset, the method comprising selecting, as a cluster attractor, the respective probability distribution of one or more terms from each subset having an entropy that satisfies the respective entropy threshold.**

(Table 1, "Threshold on influence score For example, it could be required that to further subdivide a cluster, the influence score of the highest remaining attribute had to be at least 0.05, or 0.1, etc…."), "threshold" being "assigned" as claimed.

However Sahami, does not specifically disclose: ***cluster attractors***

(Abstract and Claim 1, "…identifying an attractor for each of a plurality of clusters…").

Therefore, it would have been obvious to a person of ordinary skill in the art at the time of invention was made to incorporate the teachings of Tukey into the method of Sahami to include cluster attractors based on the probability distribution and entropy value. The modification would have been obvious because one of the ordinary skills in the art would implement an effective approach to find the best similarities using the cluster attractors and get the closest match to the comparison between the set of values to give the user the best probable result.

As per Claim 8, the rejection of Claim 7 is incorporated and further Sahami discloses:

- **comprising selecting, as a cluster attractor, the respective probability distribution of one or more terms from each subset having an entropy that is less than or equal to the respective entropy threshold.**

(Par [0093], "In this example, the institution is a local charity with most of its contributors in only one zip code. In step 500, the entropy value for the zip code attribute will be low, reflecting its non-uniformity. In this example, a predetermined threshold has been established and attributes with an entropy higher than the threshold are eliminated. In this example, the entropy for the zip code attribute within the example set of data does not exceed the threshold and zip code is will not be eliminated as an attribute.")

However Sahami, does not specifically disclose: ***cluster attractors***

O n the other hand Tukey discloses the claimed feature as follow:

(Abstract and Claim 1, "…identifying an attractor for each of a plurality of clusters…").

Therefore, it would have been obvious to a person of ordinary skill in the art at the time of invention was made to incorporate the teachings of Tukey into the method of Sahami to include cluster attractors based on the probability distribution and entropy value. The modification would have been obvious because one of the ordinary skills in the art would implement an effective approach to find the best similarities using the cluster

attractors and get the closest match to the comparison between the set of values to give the user the best probable result.

As per Claim 9, the rejection of Claim 5 is incorporated and further Sahami discloses:

- wherein each subset is associated with a frequency range and wherein the frequency ranges for respective subsets are disjoint.

(Abstract and Claim 5, "The method of claim 1, wherein the computing a frequency information for the set of attributes comprises performing a CUBE operation on the set of data for the set of attributes that **computes the frequency information**.")

However Sahami, does not specifically disclose: **for respective subsets are disjoint.**

O n the other hand Tukey discloses the above underlined claimed features as follows:

(Col. 13, lines 8-16, "This output is then the inner cluster information that is more precise than that which is obtained **either from a non-disjoint or a disjoint clustering** of all documents in the input corpus. Alternatively, since the inner clusters need not span the next-upward cluster to which they belong, intermediate clusters can be defined corresponding to each of the overlapping clusters and consisting of all documents, in that overlapping cluster, for which the corresponding attractor is the closest attractor.")

Therefore, it would have been obvious to a person of ordinary skill in the art at the time of invention was made to incorporate the teachings of Tukey into the method of Sahami to include disjoint frequency values for each subset. The modification would have been

obvious because one of the ordinary skills in the art would implement an effective approach to find the best similarities using the cluster attractors and get the closest match to the comparison between the set of values to give the user the best probable result.

As per Claim 10, the rejection of Claim 5 is incorporated and further Sahami discloses:

> **- wherein each subset is associated with a frequency range, the size of each successive frequency range being equal to a constant multiplied by the size of the preceding frequency range in order of increasing frequency.**

(Par [0076], "One enhancement that can improve the quality of the decision tree is to form subsets. Once an attribute x. sub.i has been selected based on the influence for x.sub.i, the attribute with which x.sub.i had the highest mutual information value is also known, e.g. x.sub.j. By writing out vectors for the **frequency** of the various attribute values of x.sub.j for each of the attribute values of x.sub.i, a decision can be made as to whether subsetting is appropriate.") and (Par [0009], "As the number of attributes **increases, the vector increases** in length. For n attributes, a data point X can be represented by an n-vector:").

As per Claim 11, the rejection of Claim 7 is incorporated and further Sahami discloses:

- wherein the respective entropy threshold increases for successive subsets in order of increasing frequency.

(Par [0093], “In this example, the institution is a local charity with most of its contributors in only one zip code. In step 500, the entropy value for the zip code attribute will be low, reflecting its non-uniformity. In this example, a predetermined **threshold has been established and attributes with an entropy** higher than the threshold are eliminated. In this example, the entropy for the zip code attribute within the example set of data does not exceed the threshold and zip code is will not be eliminated as an attribute.”).

As per Claim 12, the rejection of Claim 11 is incorporated and further Sahami discloses:

- wherein the respective entropy threshold for successive subsets increases linearly.

(Par [0009], “As the number of attributes **increases, the vector increases** in length. For n attributes, a data point X can be represented by an n-vector:”) and (Par [0087] In some embodiments, a predetermined value can be used to automatically eliminate attributes with higher entropies. Alternatively, the m attributes with the highest entropies can be eliminated. Alternatively, the j attributes with the lowest entropies can be the only ones used in clustering.").

As per Claim 13, Sahami discloses:

- A computer implemented **method of determining** **cluster attractors** **for a plurality of documents, each document comprising at least one term,** each term comprising one or more words**, the method comprising: calculating, in respect of each term, a probability distribution indicative of the frequency of occurrence of the, or each, other term that co-occurs with said term in at least one of said documents; calculating, in respect of each term, the entropy of the respective probability distribution; selecting at least one of said probability distributions as a** **cluster attractor** **depending on the respective entropy value.**

(Abstract, "…include methods for identifying clusters in a database, data warehouse or data mart. The identified clusters can be meaningfully understood by a list of the attributes and corresponding values for each of the clusters…") and (Par [0058], "The do not care entry is useful in probabilistic algorithms because the frequency of different attribute values is easily computed. For example, the probability that x.sub.1=a is: 1 P (x 1 = a) = (x 1 = a , x 2 = * , x 3 = *) (x 1 = * , x 2 = * , x 3 = *) ,") and (Par [0059]) and (Par [0062], "The mutual…") and (Par [0027], "The previously discussed techniques were all oriented towards clustering entire sets of data. COBWEB is an online, or incremental approach to clustering. FIG. 4 shows a COBWEB tree structure with clusters. The clusters are the nodes of the tree. FIG. 4 shows a new data point, X, to be added to the data set. COBWEB is based on **a probability distribution** …").

However Sahami, does not specifically disclose:

- ***cluster attractors***

- each term comprising one or more words

On the other hand Tukey discloses the above claimed features as follow:

(Abstract and Claim 1, "…identifying an attractor for each of a plurality of clusters…"). (Col. 6 and Col. 7, lines 65-67 and 1-9, respectively, "…A feature may be a word, a statistical phrase…").

Therefore, it would have been obvious to a person of ordinary skill in the art at the time of invention was made to incorporate the teachings of Tukey into the method of Sahami to include cluster attractors based on the probability distribution and entropy value. The modification would have been obvious because one of the ordinary skills in the art would implement an effective approach to find the best similarities using the cluster attractors and get the closest match to the comparison between the set of values to give the user the best probable result.

As per Claims 14, being the **apparatus claim** corresponding to the method 1, respectively and rejected under the same reason set forth in connection of the rejections of Claim 1, and further Sahami discloses: (Title: "METHOD AND APPARATUS FOR SCALABLE PROBABILISTIC CLUSTERING …")

As per Claim 15, the rejection of Claim 1 is incorporated and further Sahami discloses:

A method of clustering a plurality of documents, each document comprising at least one term, each term comprising one or more words, the method comprising determining cluster attractors in accordance with Claim 1.

(Par [0137], "FIG. 11 shows the output ... The remaining portion 1102 of the document (web browser display is scrolled to see all of it) is the output from the **clustering** process.").

However Sahami does not specifically teaches:

comprising at least one term, *each term comprising one or more words*

the method comprising determining cluster attractors

On the other hand Tukey discloses the above underlined claimed features as follows:

(Col. 6 and Col. 7, lines 65-67 and 1-9, respectively, "…A feature may be a word, a statistical phrase…").

And (Col. 10, lines 34-44, "…a particular cluster based upon its "closeness" to the cluster's attractor as, for example, described above. However, the present invention further modifies the closeness criterion, which would result in disjoint (non-overlapping) clusters. The modification of the clustering is accomplished by adding to each cluster at least one document found in another cluster. As illustrated by step 88, the number of additional clusters is determined as a fraction of the number of documents in each cluster.")

Therefore, it would have been obvious to a person of ordinary skill in the art at the time of invention was made to incorporate the teachings of Tukey into the method of Sahami to include cluster attractors based on the probability distribution and entropy value. The modification would have been obvious because one of the ordinary skills in the art would implement an effective approach to find the best similarities using the cluster attractors and get the closest match to the comparison between the set of values to give the user the best probable result.

As per Claim 16, the rejection of Claim 15 is incorporated and further Sahami discloses:

> **- comprising: calculating, in respect of each document, a probability distribution indicative of the frequency of occurrence of each term in the document; comparing the respective probability distribution of each document with each probability distribution selected as a cluster attractor; and assigning each document to at least one cluster depending on the similarity between the compared probability distributions.**

(Abstract, "…include methods for identifying clusters in a database, data warehouse or data mart. The identified clusters can be meaningfully understood by a list of the attributes and corresponding values for each of the clusters…") and (Par [0058], "The do not care entry is useful in probabilistic algorithms because the frequency of different attribute values is easily computed. For example, the probability that x.sub.1=a is: 1 P (

x 1 = a) = (x 1 = a , x 2 = * , x 3 = *) (x 1 = * , x 2 = * , x 3 = *) ,") and (Par [0059]) and (Par [0062], "The mutual…") and (Par [0027], "The previously discussed techniques were all oriented towards clustering entire sets of data. COBWEB is an online, or incremental approach to clustering. FIG. 4 shows a COBWEB tree structure with clusters. The clusters are the nodes of the tree. FIG. 4 shows a new data point, X, to be added to the data set. COBWEB is based on **a probability distribution** …").

However Sahami, does not specifically disclose: ***cluster attractors***

On the other hand Tukey discloses the above underlined claimed feature as follows:

cluster attractors

(Abstract and Claim 1, "…identifying an attractor for each of a plurality of clusters…").

Therefore, it would have been obvious to a person of ordinary skill in the art at the time of invention was made to incorporate the teachings of Tukey into the method of Sahami to include cluster attractors based on the probability distribution and entropy value. The modification would have been obvious because one of the ordinary skills in the art would implement an effective approach to find the best similarities using the cluster attractors and get the closest match to the comparison between the set of values to give the user the best probable result.

As per Claim 17, the rejection of Claim 16 is incorporated and further Sahami discloses:

- comprising organizing the documents within each cluster by: assigning a respective weight to each document, the value of the weight depending on the similarity between the probability distribution of the document and the probability distribution of the cluster attractor; comparing the respective probability distribution of each document in the cluster with the probability distribution of each other document in the cluster; assigning a respective weight to each pair of compared documents, the value of the weight depending on the similarity between the compared respective probability distributions of each document of the pair; calculating a minimum spanning tree for the cluster based on the respective calculated weights.

(Par [0013], "In order to position the centroids and define the clusters, the k-means technique relies on the existence of a **similarity**, or distance, function for the domain…") and (Par [0017], "…Then, at the next step, **a similarity**, or distance function is used to find the closest pair of smaller clusters, which are then merged into a larger cluster…") and (Par [0027], "The previously discussed techniques were all oriented towards clustering entire sets of data. COBWEB is an online, or incremental approach to clustering. FIG. 4 shows a COBWEB tree structure with clusters. The clusters are the nodes of the tree. FIG. 4 shows a new data point, X, to be added to the data set. COBWEB is based on **a probability distribution** …") and (Par [0039] FIG. 1 illustrates the centroids and clusters determined by the **k-means algorithm** on a data set with two attributes and k=3.") and (Par [0069] The attribute with the highest influence is then selected. The selected attribute is the one upon which the remaining attributes

are most heavily dependent.) and (Par [0070] One optimization is to sum over the k-maximal terms instead of all of the other attributes in computing the influence. In this embodiment, only the k mutual influence values with the highest values are summed for each attribute."), the k-means algorithm, being the "calculating a minimum spanning tree for the cluster".

However Sahami, does not specifically disclose:

cluster attractors

respective weight to each document

assigning a respective weight to each pair of compared documents, the value of the weight depending on the similarity

On the other hand Tukey discloses the above claimed features as follows:

(Abstract and Claim 1, "…identifying an **attractor for each of a plurality of clusters**…").

And (Cols. 6 and 7, lines 65-68 and 1-12, respectively, "partially described so as to enable a similarity determination. A document may, for example…value decomposition (SVD) analysis of the word by document matrix), or similar unit of understanding into which the document may be divided. SVD is a matrix factorization technique. Basically, **a words versus document co-occurrence** matrix is factored via SVD and only the highest weighted rotated").

Therefore, it would have been obvious to a person of ordinary skill in the art at the time of invention was made to incorporate the teachings of Tukey into the method of Sahami

to include cluster attractors based on the probability distribution and entropy value. The modification would have been obvious because one of the ordinary skills in the art would implement an effective approach to find the best similarities using the cluster attractors and get the closest match to the comparison between the set of values to give the user the best probable result.

As per Claim 18, Sahami discloses:

A computer-implemented method of clustering a plurality of documents, each document comprising at least one term, each term comprising one or more words, the method including: causing a computer to calculate, in respect of each term, a probability distribution indicative of the frequency of occurrence of the, or each, other term that co-occurs with said term in at least one of said documents; causing the computer to calculate, in respect of each term, the entropy of the respective probability distribution; causing the computer to select at least one of said probability distributions as a cluster attractor depending on the respective entropy value.

(Abstract, "…include methods for identifying clusters in a database, data warehouse or data mart. The identified clusters can be meaningfully understood by a list of the attributes and corresponding values for each of the clusters…") and (Par [0058], "The do not care entry is useful in probabilistic algorithms because the frequency of different attribute values is easily computed. For example, the probability that x.sub.1=a is: 1 P (x 1 = a) = (x 1 = a , x 2 = * , x 3 = *) (x 1 = * , x 2 = * , x 3 = *) ,") and (Par [0059]) and

(Par [0062], "The mutual…") and (Par [0027], "The previously discussed techniques were all oriented towards clustering entire sets of data. COBWEB is an online, or incremental approach to clustering. FIG. 4 shows a COBWEB tree structure with clusters. The clusters are the nodes of the tree. FIG. 4 shows a new data point, X, to be added to the data set. COBWEB is based on **a probability distribution** …").

However Sahami, does not specifically disclose:

- each term comprising one or more words

On the other hand Tukey discloses the above claimed features as follow:

(Col. 6 and Col. 7, lines 65-67 and 1-9, respectively, "…A feature may be a word, a statistical phrase…").

Therefore, it would have been obvious to a person of ordinary skill in the art at the time of invention was made to incorporate the teachings of Tukey into the method of Sahami to include clustering based on one or more words, the probability distribution and entropy value. The modification would have been obvious because one of the ordinary skills in the art would implement an effective approach to find the best similarities using the cluster attractors and get the closest match to the comparison between the set of values to give the user the best probable result.

Conclusion

10. Applicant's amendment necessitated the new ground(s) of rejection presented in this Office action. Accordingly, **THIS ACTION IS MADE FINAL**. See MPEP

§ 706.07(a). Applicant is reminded of the extension of time policy as set forth in 37 CFR 1.136(a).

A shortened statutory period for reply to this final action is set to expire THREE MONTHS from the mailing date of this action. In the event a first reply is filed within TWO MONTHS of the mailing date of this final action and the advisory action is not mailed until after the end of the THREE-MONTH shortened statutory period, then the shortened statutory period will expire on the date the advisory action is mailed, and any extension fee pursuant to 37 CFR 1.136(a) will be calculated from the mailing date of the advisory action. In no event, however, will the statutory period for reply expire later than SIX MONTHS from the date of this final action.

11. Any inquiry concerning this communication or earlier communications from the examiner should be directed to ANGELICA RUIZ whose telephone number is (571)270-3158. The examiner can normally be reached on 8:00 a.m. to 4:30 p.m., ET.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Mohammad Ali can be reached on (571) 272-4105. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

Angelica Ruiz
Examiner
Art Unit 2169

/J. M. C./

/Mohammad Ali/

Supervisory Patent Examiner, Art Unit 2169